Exploring the Use of Main Memory Database (MMDB) Technology for the Analysis of Genomic Microarray Data

Carriero, Nicholas[1], Osier, Michael[2], Cheung, Kei[2], Masiar, Peter[2], Schultz, Martin[1], Miller, Perry[*2,3,4]
[1]Department of Computer Science, [2]Center for Medical Informatics, [3]Department of Molecular, Cellular and Developmental Biology, [4]Department of Anesthesiology, Yale University, New Haven, CT, USA

We are exploring the use of main memory database (MMDB) technology in support of genomic microarray analysis, as part of a broader project exploring the use of high performance computation (HPC) in biomedicine. A great deal of work involving biomedical HPC has focused on allowing compute-intensive applications to be executed in parallel. MMDB technology represents a complementary form of HPC that allows computations that might otherwise be executed within a conventional database management system to be executed in main memory, thereby avoiding much of the overhead imposed by disk access that is implicit in conventional database systems. The use of MMDB technology is becoming increasingly feasible as the cost of main memory plummets and as an increasingly large amount of such memory is supported by individual workstations. MMDB technology can also be combined with parallel computation in parallel main memory database (PMMDB) applications.

MMDB approaches can be particularly valuable when certain types of analyses need to be performed on data stored in a database, but are not an ideal match with the type of application for which a conventional database is designed. This is particularly true, for example, if a computation involves a complex analysis that iterates over large tables of data, as is often required for the analysis of microarray data. Experimental microarray data are a challenge to archive and analyze efficiently using a relational approach. On the one hand, some of the data (the administrative and descriptive data) may lend themselves naturally to the methods of traditional relational databases. On the other hand, the high-throughput data (spot readings) are inherently tabular numeric values and the "natural" computation over these data is often more easily conceived and expressed in terms of mathematical operations on vectors and matrices, rather than in terms of Boolean operations expressed in SQL on the elements of a relational database.

We are exploring these issues in the context of the Yale Microarray Database (YMD), an institutional database designed to help support many different microarray researchers at Yale, including collaborators at other institutions. YMD currently manages microarray data sets using standard relational database technology. We have performed a variety of tests involving queries performed on YMD microarray data comparing the performance of the relational framework vs. the MMDB framework. Preliminary results show dramatic performance enhancement using the MMDB approach with speedup in the range of 40-300 times. In other words, a query that takes roughly 200 sec using a relational database runs in roughly 0.7-5 sec using MMDB. We are continuing to explore various specific aspects of the performance of the two approaches, and plan to extend the work in the future to explore parallel MMDB as well.

This research is supported in part by NIH grants P20 LM07253 and T15 LM07056 from the National Library of Medicine, by NIH grants R24 DK58776 and K25 HG02378, and by NSF grant DBI-0135442.